Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 7 (2011) 246–251

Procedia Environmental Sciences

Spatial Statistics 2011

Geostatistics for spatial extremes. A case study of maximum annual rainfall in Portugal**

Manuela Neves[a*], Dora Prata Gomes[b] a*

[a]*ISA-UTL and CEAUL, Tapada da Ajuda, 1349-017 Lisboa, Portugal*
[a]*FCT-UNL and CMA, Monte de Caparica, 2829-516 Caparica, Portugal*

Abstract

Spatial statistics deals with statistical methods in which spatial locations play an explicit role in the analysis of data. Spatial data are often modeled as a realization of a Gaussian process or a function of a Gaussian process. However there are events, such as rain, snow, storms, hurricanes, earthquakes, where extremes are of the main interest. Here multivariate normal distributions are inappropriate for modeling tail behavior. The natural class of processes for dealing with extremes is the max-stable processes. This work briefly reviews the approaches to the statistical modeling of spatial extremes. Models for max-stable processes will be considered in an application to the annual maxima of daily precipitation over the North of Portugal.

Keywords: Geostatistics; Spatial extremes; max-stable processes; extreme precipitation; modeling

1. Introduction

Spatial data contain locational information as well as attribute information, i.e., they are data for which some attribute are recorded at different locations and these locations are coded as part of the data. Spatial analysis is a general term to describe a technique that uses that information in order to better understand

*.Corresponding author. Tel.: +365213653325; fax: +365213630723.
E-mail address: manela@isa.utl.pt.
**Research partially supported by FCT/OE, POCI2010 and PTDC/FEDER.

doi:10.1016/j.proenv.2011.07.043

the processes generating the observed attribute. The general spatial model is that one that describes the characteristics of $Y(\boldsymbol{x})$, where Y is a random outcome variable occurring at a point **x** in n-space. Spatial data and the processes generating such data have several properties that distinguish them from other data. Firstly, the data are typically not independent of each other. Attribute values in nearby places tend to be more similar than are attribute values drawn from locations far away from each other. This is a useful property when it comes to predicting unknown values because we can use the information that an unknown attribute value is likely to be similar to neighboring, known values. The subfield of geostatistics has grown up based on this premise. However, spatial autocorrelation causes problems for those statistical techniques that assume data drawn from independent random samples. Special statistics methods have been developed to overcome this problem.

Historically, Gaussian processes play a central role in modeling spatial processes. Spatial data are often modeled as a realization from a Gaussian process or a function of a Gaussian process. However there are events, such as rain, snow, storms, hurricanes, earthquakes, where extremes are of main interest. Here multivariate normal distributions are inappropriate for modeling tail behavior. The most natural way for the continuous space specification of extremes is provided by the theory of max-stable processes, which can be seen as an extreme value analogy of Gaussian processes. Two different characterizations have been proposed in [1] and [2]. This work reviews briefly approaches to the statistical modeling of spatial extremes. Geostatistical max-stable models will be fitted to an application to the annual maxima of daily precipitation over the North of Portugal. First steps in the exploratory analysis of those data have already been presented in [3], but here we intend to explore also recent developments in R software [4].

2. Stochastic Models for spatial data

A spatial process is a stochastic process, represented by a set of random variables (or vectors) $\boldsymbol{Y}(\boldsymbol{x})$, indexed over $\boldsymbol{x}$ defined on a set $D \subset IR^d$, a d-dimensional Euclidean space. Usually d = 1,2,3. The usual notation is $\{\boldsymbol{Y(x)}: \boldsymbol{x} \in D\}$. Observations are taken in spatial locations $S = \{\boldsymbol{x}_1,...,\boldsymbol{x}_n\}$ and can be from one or more discrete or continuous variables. The nature of D allows us to identify three types of data and consequently three main spatial processes: point-referenced data, areal data and point pattern data, see [5]. Here we mention only point-referenced data where S is finite but $\boldsymbol{x}$ can vary continuously in D. The phenomenon could be observed in all fixed points of this domain and therefore can be modeled by a spatial process, $\{Y(\boldsymbol{x}): \boldsymbol{x} \in D\}$ and D is a subset of IR^2 containing a rectangle of area non-zero. The purpose of the study is to predict the value of the feature analyzed in points not observed, i.e. to extrapolate from S to the whole D. Here we will only refer to cases where observations $y(\boldsymbol{x})$ are one-dimensional and the domain is in IR^2.

Models used in spatial data need to incorporate the correlation structure in order to increase the accuracy of modeling and prediction efforts. The dependence between observations can be translated into a function of distance between locations. Essential elements of point-referenced data for modeling are stationarity, isotropy and the variogram, the key elements of the "Matheron school". By stationarity we mean that the distribution of the random process has attributes that are the same everywhere. The stationarity is a property that we need to consider for inference. The more commonly considered is weak stationarity that occurs if the covariance relationship between the values of the process at any two locations can be summarized by the covariance function, $C(\boldsymbol{h})$, sometimes called covariogram, that depends only on the separation vector $\boldsymbol{h}$.

Suppose we assume $E[Y(\boldsymbol{x}+\boldsymbol{h})-Y(\boldsymbol{x})]=0$ and define $E[Y(\boldsymbol{x}+\boldsymbol{h})-Y(\boldsymbol{x})]^2=\mathrm{Var}[Y(\boldsymbol{x}+\boldsymbol{h})-Y(\boldsymbol{x})]=2\gamma(\boldsymbol{h})$. The quantity $2\gamma(\boldsymbol{h})$ is named the variogram, and $\gamma(\boldsymbol{h})$ is the semivariogram. In above expression we are saying that $\gamma(\boldsymbol{h})$ do not depend on $\boldsymbol{x}$**;** the process is said to be intrinsically stationary. If the semivariogram depends upon the separation vector only through its length $||\boldsymbol{h}||$, $(\gamma(||\boldsymbol{h}||))$, the process is said to be

isotropic. If the process is intrinsically stationary and isotropic, it is also called homogeneous. For this situation, several models for the variogram have been extensively studied, see [5] and [6].

The process $\{Y(\boldsymbol{x}): \boldsymbol{x}\in D\}$ is said to be Gaussian if for any $n \geq 1$ and locations $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$, the vector $(Y(\boldsymbol{x}_1),\ldots,Y(\boldsymbol{x}_n))$ has a multivariate normal distribution. It is well known that if $Y(.)$ is a stationary Gaussian process with correlation function ρ and variance σ^2 it is completely characterized and the variogram is defined as $\gamma(\boldsymbol{x}_1-\boldsymbol{x}_2)= (1/2)\mathrm{Var}[Y(\boldsymbol{x}_1)-Y(\boldsymbol{x}_2)]=\sigma^2(1-\rho(\boldsymbol{x}_1-\boldsymbol{x}_2))$.

In a practical situation the procedure for geostatistical inference is done in several steps: remove trends in mean and (perhaps) variance of data; transform residuals to standard normal margins; use graphical techniques to assess likely form of covariance function, anisotropy, etc.; fit suitable covariance model; make inferences using kriging, likelihood or Bayesian procedures; make predictions using the fitted correlation function and then add back the estimated trends to obtain a map of predictions, based on fitted normal model. Geostatistics is well developed but is largely based on multivariate normal distributions, inappropriate to model the tail of a distribution.

3. Geostatistics for spatial extreme data

Many environmental extremal problems are spatial or temporal in nature, or both, such as precipitation, storms, avalanches. Application of extreme value methods to environmental processes has long been established as one of the main practical uses of the extreme value theory. Most often univariate methods have been used to describe the extremal behavior of a process at a given location. Occasionally, the joint behavior of extreme values of the processes is inherently spatial and for some processes there is interest in features of the extreme spatial events. Despite this, there have been only a very few applications which exploit the spatial structure of the extreme values of the process, see for example [7] and [8] for rainfalls, [9] and [10] for sea-levels. There is little guidance for the modeling of spatial extremes.

Understanding the spatial-temporal variability of these extreme events is crucial to predict extreme events. However conventional geostatistics are not relevant here because extremes are far from being normal and variogram based approaches may not even exists, because it may happen $\mathrm{E}[Y(\boldsymbol{x})]= +\infty$ or $\mathrm{Var}[Y(\boldsymbol{x})] = +\infty$. Extending the Extreme Value Theory to the spatial case is a recent research area. [1] suggests a procedure using the theory of max-stable processes for modeling data, which are collected on a grid of points in space. As it is well known, fundamental to all characterizations of extreme value processes is the concept of stability. The generalized extreme value distribution (GEV) is used to model scalar extremes because of its max-stability, which gives a mathematical basis for extrapolation beyond the range of the data.

3.1. Max-stable processes and dependence measures

The natural models for spatial extremes are max-stable processes which extend the GEV to spatial data. A max-stable process $Z(.)$ is the limit process of maxima of i.i.d. random fields $Y_i(\boldsymbol{x})$, $\boldsymbol{x}\in IR^d$, i.e., for suitable $a_n(\boldsymbol{x}) > 0$ and $b_n(\boldsymbol{x})\in IR$,

$$\lim_{n\to\infty}\{(\max_{i=1,\ldots,n}Y_i(\boldsymbol{x}) - b_n(\boldsymbol{x})) / a_n(\boldsymbol{x})\} =Z(\boldsymbol{x}), \quad \boldsymbol{x}\in IR^d \qquad (1)$$

Provided limit in (1) exists the max-stable processes are appropriate to model annual maxima of spatial data, for example. Max-stable processes are used by [11] to model directional dependence of extreme wind speeds and [8] to develop a model for spatially aggregated rainfall extremes. Without losing generality, is more convenient for practical purposes, margins can be transformed to one particular

extreme value distribution and it turns out to be convenient to assume the standard Fréchet distribution, $P(Z(\boldsymbol{x}) \leq z) = \exp\{-1/z\}$, for all $\boldsymbol{x} \in IR^d$ and $z > 0$. [12] introduced a very useful representation of max-stable processes, that allowed [1] to provide a parametric model for spatial extremes. More recently [2] introduced a second characterization of max-stable processes.

Such as in the classical approach, to know how evolves the dependence in space is fundamental for modeling spatial extremes data. For classical geostatistics the common tools is the (semi-)variogram. For max-stable processes, to study how decreases the dependence between two locations when the distance increases, is done by some measures of dependence. Those measures are the **extremal coefficient function** and some **variogram-based approach** designed for extremes.

The extremal dependence among n fixed locations in IR^d can be summarized by the extremal coefficient, defined as $P(Z(\boldsymbol{x}_1) \leq z,...,Z(\boldsymbol{x}_n) \leq z) = \exp\{-\theta_n/z\}$, where $1 \leq \theta_n \leq n$, can be thought as the effective number of independent location: $\theta_n = 1$, perfect dependence and $\theta_n = n$, independence. An important special case of that equation is to consider pairwise extremal coefficients, i.e.,

$$P(Z(\boldsymbol{x}_1) \leq z\ ,\ Z(\boldsymbol{x}_2) \leq z] = \exp\{-(\theta(\boldsymbol{x}_1) - \theta(\boldsymbol{x}_2))/z\} \qquad (2)$$

For Smith and Schlatter models, we have, respectively $\theta(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 2\Phi(\sqrt{((\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2))}/2)$ and $\theta(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 1 + \sqrt{(1-\rho(\boldsymbol{x}_1 - \boldsymbol{x}_2))}/2)$. Estimators for extremal coefficient are the maximum likelihood estimator of Schlather and Tawn [13] and an estimator of Smith [1].

When we are dealing with extremes the variogram may not exist, therefore there is a need of developing suitable tools for analyzing the spatial dependence of max-stable fields. A standard tool, similar to the variogram, is the Madogram [14], defined as $\upsilon(\boldsymbol{x}_1 - \boldsymbol{x}_2) = E[|Z(\boldsymbol{x}_1) - Z(\boldsymbol{x}_2)|]$. As estimators for Madogram, there are pairwise estimator, the binned pairwise estimator (if isotropy) and, by relation between pairwise extremal coefficient and the Madogram, the plug-in estimator for pairwise extremal coefficient. The F-Madogram, see [15], is defined as $\upsilon_F(\boldsymbol{x}_1 - \boldsymbol{x}_2) = E[|F\{Z(\boldsymbol{x}_1)\} - F\{Z(\boldsymbol{x}_2)\}|]$. As estimators for F-Madogram there are the binned estimator, by plugging an estimate of the CDF at the specified location, and by relation between pairwise extremal coefficient and the F-Madogram the plug-in estimator for pairwise extremal coefficient.

The λ-Madogram [16] defined as $\upsilon_\lambda(\boldsymbol{x}_1 - \boldsymbol{x}_2) = (1/2)E(|F^\lambda\{Z(\boldsymbol{x}_1)\} - F^{1-\lambda}\{Z(\boldsymbol{x}_2)\}|)$. The F-Madogram is similar to λ-Madogram when $\lambda = 0.5$. Estimators for λ-Madogram there are the binned λ-Madogram estimator and the adjusted estimator, see [15].

4. Application to maximum annual precipitation data in North of Portugal

We now consider an application of the procedures referred to above to the annual maximum values of daily rainfall in the North of Portugal. At 21 different sites and for 20 hydrological years (1977-1996) the annual maxima of daily precipitations were recorded. The first steps in the statistical study of those data were done in [3].

Here we intend to analyze the spatial behavior of the data and to fit max-stable models in order to get high precipitation level maps. Our study was done in R statistical computing program (http://cran.r-project.org/) an enormously successful open-source system on statistical computing and modeling.

Several packages on Extreme Value analysis have been recently introduced into R, but more recently Ribatet added to R the *SpatialExtremes* that provides functions to analyze and fit max-stable processes to spatial extremes.

Fig.1(a) shows the map with locations of several stations in North of Portugal and Fig. 1(b) shows the positions of the stations from which the data have been recorded (for other stations, there exist many missing values in the series of observations – a challenge for future research).

Some locations showed a strong departure from the normal. Fig. 2 shows the qqnormal plot for four of those locations. The GEV is fitted to data at each site. Fig. 3 shows a graphical diagnostic for the GEV fitting in one of the locations. Data are transformed at each site so that they have a standard Fréchet distribution.

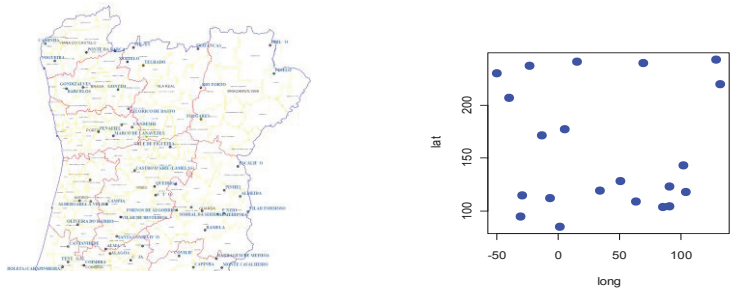

Fig. 1. (a) Map of locations of several stations in the North of Portugal. (b) Positions of our data.

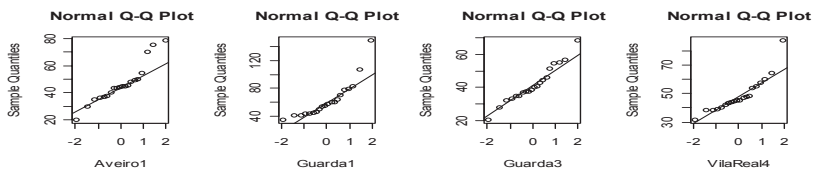

Fig. 2. The qqnormal plot for some locations.

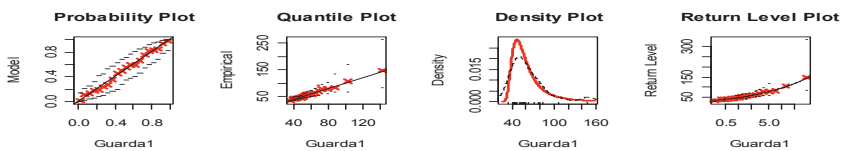

Fig 3. GEV model diagnostic at one of the locations (Guarda1).

For estimating the spatial dependence structure , the extremal coefficient between each pair of sites is calculated, according to Smith and Schlatter-Tawn models. Fig. 4 shows the pairwise extremal coefficient estimates and lowess curves (left). The plots show high variability. Fig. 4 (right) shows the F-madogram and the binned F-madogram. The distance between locations shows almost independence when locations are far from 120/150km .

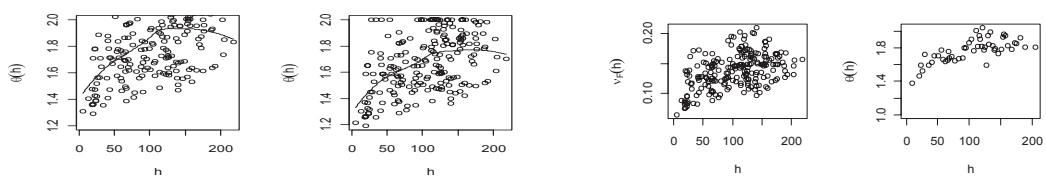

Fig 4. Extremal coefficient estimates against the distance h; (a) Smith model; (b) Schlatter and Tawn model (left). F-madogram and binned F-madogram (on the right).

Fig. 5 shows the generate max-stable random field in the region under study, using the estimated parameter of Smith model, with Gaussian correlation and Schalter model with power correlation.

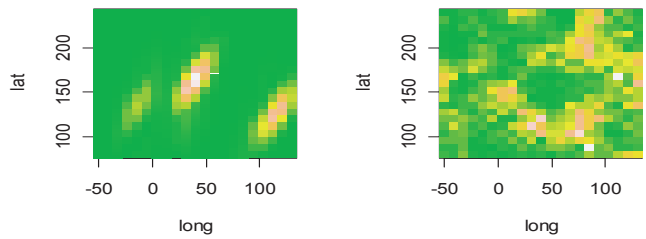

Fig 5. Output of the generate max-stable random field using estimated parameteres of Smith model, with Gaussian correlation (a) and Schalter model with power correlation (b).

References

[1] Smith R. Max-stable processes and spatial extremes. *Unpublished manuscript* 1990.

[2] Schlather M. Models for stationary max-stable random fields. *Extremes* 2002; **5**(1): 33-44.

[3] Prata Gomes D, Neves M. Preliminary Spatial study of high precipitation levels in North of Portugal. *Workshop on Statistical Aspects of Environmental Risk* 2010, SAMSI, Carolina do Norte.

[4] Ribatet M. A user's guide to the SpatialExtremes package 2009. *Available as a part of the package or at the package web page*. École Polytechnique Fédérale de Lausanne, Switzerland.

[5] Cressie NA. *Statistics for spatial data* 1991. Wiley, New York.

[6] Cressie N, Hawkins D M. Robust estimation of the variogram. *Journal of the International Association for Mathematical Geology* 1980;**12**:115-125.

[7] Coles S G. Regional modeling of extreme storms via max-stable processes. Journal of the Royal Statistical Society 1993; **55**(B):797-816.

[8] Coles S G, Tawn J A. Modeling extremes of the areal rainfall process. Journal of the Royal Statistical Society 1996; **58**(B):329-347.

[9] Coles S G, Tawn J A. Statistics of coastal flood prevention. *Philosophical Transactions of the Royal Society* 1990; **332**(A): 457-476.

[10] Dixon J M, Tawn J A, Vassie J M. Spatial modeling oextreme sea-levels. *Environmetrics* 1998;**9**(3):283-301.

[11] Coles S, Walshaw D. Directional Modeling of extreme wind speeds. *Journal of the Royal Statistical Society* C 1994;**43**(1):139-157.

[12] Haan L de. A spectral representation for max-stable processes. *The Annals of Probability* 1984; **12**(4): 1194-1204.

[13] Schlather M, Tawn J A. A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika* 2003;**90**(1):139–156.

[14] Matheron G. Suffit-il, pour une covariance, d'eˆtre de type positif?: *Sciences de la Terre*, *série informatique géologique* 1987; **26**:51–66.

[15] Naveu Cooley D, Naveau P, Poncet P. Variograms for spatial max-stable random fields. In: Springer, editor, *Dependence in Probability and Statistics* Springer, New York, Lecture notes in statistics edition 2006;187:373-390.

[16] Naveau P, Guillou A, D Cooley, J Diebolt. Modeling Pairwise Dependence of Maxima in Space. *Biometrika* 2009; 96(1):1–17.